# A unified approach to word occurrence probabilities

Mireille Regnier

# A Unified Approach to Word Occurrence Probabilities

Mireille Régnier

*INRIA, 78153 Le Chesnay, France*

**Abstract**

Evaluation of the expected frequency of occurrences of a given set of patterns in a DNA sequence has numerous applications and has been extensively studied recently. We provide a unified framework for this evaluation that adapts to various constraints and allow to extend previous results. We assume successively that the patterns may, then may not, overlap. We derive exact formulae for the moments in a Markovian model, that are linear functions of the size of the sequence. We show that our formulae, that occasionnally simplify previous results, are computable at low cost, which makes them useful for practical applications.

## 1 Introduction

Repeated patterns and related phenomena in sequences (also called words or strings) are studied in molecular biology. A survey on various methods can be found in [21]. One fundamental question that arises is the frequency of pattern occurrences in another string known as the *text*. This question is addressed below for a set of patterns ($H_i$) and various assumptions on the counting of possible overlaps. The text may be generated according either to the Bernoulli model or the Markovian model. Among the problems of molecular biology that can benefit from these results, one may cite the search of patterns with unexpectedly high or low frequencies [14] and gene recognition based on statistical properties [31,12]. Statistical methods have been successfully used from the early 80's to extract information from sequences of DNA. In particular, identifying deviant short motifs, the frequency of which is either too high or too low, might point out unknown biological information [8,7,23,18]. From this perspective, these results give estimates for the statistical significance of deviations of word occurrences from the expected values and allow a

[1] This research was supported by ESPRIT LTR Project No. 20244 (ALCOM IT).

biologist to build a dictionary of contrast words in genetic texts. They have been recently used to detect dos-DNA in the yeast chromosome [10]. Another biological problem for which such results might be useful is gene recognition. Most gene recognition techniques rely on the observation that the statistics of patterns (motifs/codon) usage in coding and non-coding regions are different [9,34]. These findings allow the estimation of the statistical significance of such differences, and the construction of the confidence interval for pattern occurrences.

The problem of pattern occurrences in a random string is a classical one, [11,17,20,6,4,15,5,16,26,30,33]. In this paper, frequency of pattern occurrences is fully characterized. It is known [2,32] that the limiting distribution is "usually normal". Let us mention that large deviation results hold [28]. Results below allow an easy computation of all moments, using for instance a symbolic computation system. Additionally, derivation is not restricted to the asymptotically dominating term, usually linear, but provides final results that are exact, or up to an exponentially decreasing term, with the same computational effort. The computation of the probability of occurrences in the finite range also follows. Moreover, most parameters of interest (average number of occurrences, waiting time for the first occurrence, $r$-scans,...) steadily follow.

The method of analysis treats uniformly two probability models, Bernoulli and Markov, and various constraints on the possible overlaps of the strings. It relies on classical combinatorial methods briefly presented in the last section. It allows for a simplification of existing formulae [22,32] as well as some corrections and, moreover, provides formulae that are *computable*. E.g. the computational complexity is low and the formulae translate into algorithms that are numerically stable. As a matter of fact, some are implemented in software COMBSTRUCT. This is crucial to applications.

## 2 Basic Tools

### *2.1 Overlapping and renewal models*

Let us consider a text string $S = t_1 t_2 \dots t_n$ of length $n$ and a set $\mathcal{H}$ of patterns $(\mathrm{H_i})_{\mathrm{i=1\cdots q}}$ of lengths $(m_i)_{i=1\cdots q}$ over an alphabet $\mathcal{S}$ of size $V$. In order to ensure an unambiguous counting, one assumes that $\mathcal{H}$ is a *reduced* set of patterns [17]; e.g. no pattern in $\mathcal{H}$ is a substring of any other pattern in $\mathcal{H}$. When $\mathcal{H}$ patterns are searched in text $S$, various constraints can be imposed on the counting of overlapping occurrences. In the various models so-defined, the occurrence of a pattern from $\mathcal{H}$ that satisfies the pre-imposed constraints is called a *valid* occurrence.

In the *overlapping model*, any occurrence is valid. Notably, two overlapping patterns both contribute to the count. For example, let

$$S = AATTATTATATTATTTT$$

with $(H_1, H_2) = (TTA, TAT)$. Patterns $H_1$ and $H_2$ occur at positions $3, 6, 11$ and $4, 7, 9, 12$. All these occurrences are valid. This is the general scheme in the search of words that occur with unexpectedly high or low frequencies. A possible application is notably the search of *tandem repeats* [1,3]. The problem has been extensively studied in [2,25,28,29,22,33].

In the *renewal model*, studied in [6,32], two overlapping occurrences cannot be valid simultaneously. In a chain of overlapping occurrences, the first occurrence is always valid. An other occurrence is valid iff it does not overlap on the left with a valid occurrence. In the example above, valid occurrences of $H_1$ and $H_2$ are found at positions $3, 6$ and $9$. This is the assumption in the enzyme restriction problem. Intuitively, when one enzyme has cut on one occurrence of pattern $H_i$, an overlapping occurrence of pattern $H_j$ does not allow enzyme $j$ to be active.

Many other constraints can be chosen that define other variants. For example, one may count overlapping occurrences of *different* patterns. Or one can force a *minimal* distance between valid occurrences.

### *2.2 Probabilistic models and notations*

Throughout this paper, the pattern set is *fixed* and given, while the text string is random. More precisely, text generation follows either one of the two probabilistic models:

(B) Bernoulli Model
The text is generated randomly by a memoryless source. Every symbol $s$ of a finite alphabet is created independently of the other symbols, with probability $p_s$. The model is *uniform* if all these probabilities are equal, otherwise it is *biased*.

(M) Markovian Model
The text is a realization of a *stationary* Markov sequence of order $K$, that is, probability of the next symbol occurrence depends on the $K$ previous symbols.

Below, $P(w)$ is the *stationary probability* that the word $w$ occurs in the random text $S$ between symbols $k$ and $k + |w| - 1$ and $P(w_1|w_2)$ is the *conditional probability* that $w_1$ occurs at position $k$ knowing that $w_2$ occurs at position

$k - |w_2|$.

We adopt the following convention to work with matrices and vectors. Bold upper-case letters are reserved for vectors which are assumed to be column vectors; e.g. $\mathbf{1_q}$ denotes the unit vector with $q$ rows. The upper index "$t$" denotes transpose and $\mathbf{1_q}$ can be rewritten $(1,\ldots,1)^t$. We shall use blackboard bold letters for matrices. In particular, we write $\mathbb{I}$ for the identity matrix. $\mathbb{M}_{i,j} = m_{ij}$ denotes element with index $(i,j)$ from matrix $\mathbb{M}$ while $\mathbb{M}_i$ denotes the matrix derived from $\mathbb{M}$ by a substitution of 0s in all columns but the $i$-th one.

Below, most derivations for the Markov model deal only with the first order Markov chain ($K = 1$). One makes use of the transition matrix $\mathbb{P} = \{p_{i,j}\}_{i,j\in\mathcal{S}}$ where $p_{i,j} = \Pr\{t_{k+1} = j | t_k = i\}$. Vector $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_V)$ denotes the stationary distribution satisfying $\boldsymbol{\pi}\mathbb{P} = \boldsymbol{\pi}$, and $\Pi$ is the stationary matrix that consists of $V$ identical rows equal to $\boldsymbol{\pi}$. Finally, $\mathbb{Z}$ is the **fundamental matrix** $\mathbb{Z} = (\mathbb{I} - (\mathbb{P} - \Pi))^{-1}$ where $\mathbb{I}$ is the identity matrix.

### *2.3 Overlapping and correlation sets*

The goal of this paper is to calculate the expected frequency of multiple pattern occurrences in the text assuming either the Bernoulli or the Markovian model. It turns out that several properties of pattern occurrences depend on the so called *correlation polynomial* introduced in [17] for the Bernoulli model, extended below to the Markov model.

**Definition 1** *Given two strings* H *and* F*, the* overlapping set *of* (H, F) *is the set of* H*-suffixes that are* F*-prefixes.* F*-suffixes of the associated* F*-factorisations form the* correlation set $\mathcal{A}_{\mathrm{H,F}}$*. One defines the* correlation polynomial *of H and F as:*

$$A_{\mathrm{H,F}}(z) = \sum_{w\in\mathcal{A}_{\mathrm{H,F}}} P(w|\mathrm{H})z^{|w|}$$

*When H is equal to F,* $\mathcal{A}_{\mathrm{H,H}}$ *is named the* autocorrelation set *and denoted* $\mathcal{A}_{\mathrm{H}}$*; empty word* $\epsilon$ *is in* $\mathcal{A}_{\mathrm{H}}$*. The* autocorrelation polynomial *is defined as:*

$$A_{\mathrm{H}}(z) = \sum_{w\in\mathcal{A}_{\mathrm{H}}} P(w|H)z^{|w|} \ .$$

Intuitively, a word in $\mathcal{A}_{\mathrm{H,F}}$, when concatenated to H, creates an (overlapping) occurrence of F. For example, let H = 11011 and $F$ = 1110 be two strings over binary alphabeet $\{0,1\}$. Then $\mathcal{A}_{\mathrm{H,F}} = \{10, 110\}$ and $\mathcal{A}_{\mathrm{F,H}} = \{11\} \neq$

$\mathcal{A}_{\mathrm{H,F}}$. The associated correlation polynomials are, in biased Bernoulli model where $(p_0, p_1) = (1/3, 2/3)$, $A_{11011,1110}(z) = \frac{2}{9}z^2 + \frac{4}{27}z^3$ while $A_{1110,11011}(z) = \frac{4}{A_{\mathrm{H,F}}(z)9}z^2$. The autocorrelation polynomials are: $A_{1110}(z) = 1$ and $A_{11011}(z) = 1 + \frac{4}{27}z^3 + \frac{8}{81}z^4$. As empty word $\epsilon$ is in $\mathcal{A}_{\mathrm{H}}$ but not in $\mathcal{A}_{\mathrm{H,F}}$, the constant term of $A_{\mathrm{H}}(z)$ is always 1 while the constant term of $A_{\mathrm{H,F}}(z)$ is always 0.

Assume now that H = CGC over alphabet $\mathcal{S} = \{A, C, G, T\}$. Observe that $\mathcal{A}_{\mathrm{H,H}} = \{\epsilon, GC\}$, where $\epsilon$ is the empty word. Thus, for the uniform Bernoulli model (all symbols occur with the same probability equal to 0.25), $A_{CGC}(z) = 1 + \frac{z^2}{16}$. In the Markovian model of order one, only the last letter in the common prefix is taken into account, and one has: $A_{CGC}(z) = 1 + p_{C,G}p_{G,C}z^2$.

**Notation:** In the following, $\mathbb{A}(z)$ denotes the $q \times q$ matrix of correlation polynomials. For the given set $\mathcal{H} = (\mathrm{H_i})_{\mathrm{i=1...q}}$ of searched patterns, $\mathbb{A}(z) = (A_{\mathrm{H_i,H_j}}(z))_{i,j=1..q}$.

## 3 Language Counting

One approach to word statistics is the study of texts that contain a finite number, say $r$, of occurrences of $\mathcal{H}$ patterns. For a given $r$, this set of texts is a language -e.g. a collection of words satisfying some properties- that is denoted $\mathcal{L}_r$. This section is devoted to the combinatorial properties of such languages. The approach is rather classical in combinatorics [13]: the structure, here a language, is decomposed into smaller substructures, here sub-languages, that are already known or more easily studied.

In combinatorics on words, two basic laws of decomposition naturally arise. A language can be decomposed into the disjoint union of smaller sub-languages while the *concatenation* of words, denoted by symbol $\cdot$, defines a product on languages. More precisely, given two languages $\mathcal{A}$ and $\mathcal{B}$, their product, denoted $\mathcal{A} \cdot \mathcal{B}$ or $\mathcal{AB}$, is the set $\{a \cdot b; a \in \mathcal{A}, b \in \mathcal{B}\}$ where $a \cdot b$ is the concatenation of strings $a$ and $b$. One denotes $\mathcal{A}^+$ the set of words formed with a concatenation of a finite number of words in $\mathcal{A}$. One denotes $\mathcal{A}^* = \mathcal{A}^+ \cup \{\epsilon\}$.

It is shown below that languages $\mathcal{L}_r$ can be decomposed, using such laws, onto basic languages that satisfy some simple equations, stated below.

### *3.1 Basic languages*

Basic languages that appear relevant to word statistics are given below:

**Definition 2** *Let $\mathcal{H}$ be a set of patterns. Given a pattern* $\mathrm{H_i}$*, the* first occurrence language $\mathcal{R}_i$ *is the set of words that admit* $\mathrm{H_i}$ *as a suffix, and contain no other pattern. One denotes* $\mathcal{R}_{\mathcal{H}} = \cup_i \mathcal{R}_i$.

*The* tail language $\mathcal{U}_i$ *is the set of words $w$ such that* $\mathrm{H_i}$ *is the only valid occurrence in* $\mathrm{H_i}$w. *It contains the empty string.*

*The* minimal languages $\mathcal{M}_{i,j}$*are defined, for two patterns* $\mathrm{H_i}$ *and* $\mathrm{H_j}$ *in* $\mathcal{H}$ *as:*

*(i)* $\mathrm{H_j}$*is a suffix of* $\mathrm{H_i}$w*;*
*(ii)* $\mathrm{H_j}$ *is valid when* $\mathrm{H_i}$ *is valid;*
*(iii)* $\mathrm{H_i}$ *and* $\mathrm{H_j}$ *are the only valid occurrences in* $\mathrm{H_i}$w.

*The* $k$-minimal language $\mathcal{M}_{i,j}^{(k)}$ *is the set of words $w$ such that:*

*(i)* $\mathcal{M}_{i,j}^{(1)} = \mathcal{M}_{i,j}$ ;
*(ii)* $\mathcal{M}_{i,j}^{(k)} = \sum_{l=1}^{q} \mathcal{M}_{i,l}^{(k-1)} \mathcal{M}_{l,j},\ k \geq 2.$

Intuitively, a word is in $\mathcal{M}_{i,j}$ (respectively $\mathcal{M}_{i,j}^{(k)}$) if its concatenation to $\mathrm{H_i}$ creates one valid occurrence of $\mathrm{H_j}$ as a suffix of $\mathrm{H_i}$w (respectively creates $k$ valid $\mathcal{H}$-occurrences, the last one being $\mathrm{H_j}$, occurring as a suffix). Languages $\mathcal{M}_{i,j}^{(k)}$ are said minimal as no prefix of any word $w$ in $\mathcal{M}_{i,j}^{(k)}$ can be in $\cup_l \mathcal{M}_{i,l}^{(k)}$. It is worth noticing that matrix $(\mathcal{M}_{i,j}^{(k)})$ from Definition 2 is equal to matrix $(\mathcal{M}_{i,j})^k$.

**Remark 3** *In the renewal scheme, $\mathcal{M}_{i,j} = \mathcal{R}_j$. This equation also holds in the overlapping case whenever* $\mathrm{H_i}$ *and* $\mathrm{H_j}$ *do not overlap.*

**Example 4** *When $\mathcal{H}$ reduces to a single pattern $H = 01$, then* $\mathcal{R}_{01} = \{1\}^* \cdot \{0\}^+ \cdot \{1\} = \{1\}^* \cdot \{0\}^* \cdot \{01\}$ *and* $\mathcal{U}_{01} = \{1\}^* \{0\}^*$. *As* H *is not self-overlapping, in both counting models,* $\mathcal{M}_{01,01} = \mathcal{R}_{01}$*; moreover,* $\mathcal{M}_{01,01}^{(k)} = \mathcal{M}_{01,01}^{k}$ .

It follows from the definition that for *all constraints* the general equation below holds:

$$\mathcal{L}_r = (\cdots, \mathcal{R}_i, \cdots)^t \times (\mathcal{M}_{i,j})^{r-1} \times \begin{pmatrix} \cdots \\ \mathcal{U}_i \\ \cdots \end{pmatrix} \tag{1}$$

**PROOF.** A word in $\mathcal{L}_r$ is associated to a set of $r$ occurrences in $\mathcal{H}$: $\mathrm{H_{i_1}}, \ldots, \mathrm{H_{i_r}}$. Hence, it can be rewritten: $w_{i_1} w_{i_2} \ldots w_{i_r} u$ where $w_{i_1}$ is in $\mathcal{R}_{i_1}$ and $w_{i_j}, 2 \leq j \leq r$ is in $\mathcal{M}_{i_{j-1},i_j}$ and $u$ is in $\mathcal{U}_{i_r}$.

A major consequence of (1) is that $\mathcal{L}_r$ is fully known when first occurrence languages, minimal languages and tail languages are known. The characterisation of these languages is the goal of the next section.

### *3.2 Set equations on basic languages*

**Notation:** Let $\mathcal{W}$ denote the language of all words on a given alphabet $\mathcal{S}$.

**Proposition 5** *The tail languages satisfy the set of equations:*

$$\forall i : \mathcal{U}_i = \mathcal{W} - \sum_{j=1}^{q} \mathcal{M}_{i,j}\mathcal{W} \ . \tag{2}$$

This result follows from a simple remark: for any word $w$ in $\mathcal{W} - \mathcal{U}_i$, $\mathrm{H_i w}$ contains at least two words from $\mathcal{H}$, e.g. exists $j$ such that $w$ has a prefix in $\mathcal{M}_{i,j}$.

Now, observe that the first occurrence or initial languages definition does not depend on the model; hence, they satisfy the same equations in the overlapping and renewal model:

**Proposition 6** *The initial languages satisfy the following equations:*

$$\forall j : \mathcal{W}\mathrm{H_j} = \sum_{\mathrm{i=1}}^{\mathrm{q}} \mathcal{R}_\mathrm{i}(\mathcal{A}_{\mathrm{i,j}} + \mathcal{W}\mathrm{H_j}) \ . \tag{3}$$

**PROOF.** For any word $w$ in $\mathcal{W}\mathrm{H_j}$, the set of $\mathcal{H}$ occurrences is not empty as it contains its suffix $\mathrm{H_j}$. Assume the first $\mathcal{H}$ occurrence is $\mathrm{H_i}$. Then some word in $\mathcal{R}_i$, say $r_i$, is a prefix of $w$. Now, suffix $\mathrm{H_j}$ of $w$ may overlap $r_i$: in that case $w = r_i \cdot a_{i,j}$ where $a_{i,j} \in \mathcal{A}_{i,j}$. Remark that $a_{i,j}$ may be the empty string: this occurs if $w$ is in $\mathcal{R}_j$. Otherwise, $w = r_i \cdot t \cdot \mathrm{H_j}$ where $t$ is any string over alphabet $\mathcal{S}$.

An alternative proof relies on the remark that $\{\mathcal{R}_i\}$ can also be derived as a function of $\{\mathcal{M}_{i,j}\}$. Then, the two counting models are treated differently. In the overlapping scheme, languages $\{\mathcal{R}_i\}$ satisfy:

$$\forall i : \mathcal{R}_i = \mathcal{W}.\mathrm{H_i} - \sum_{\mathrm{j}} \mathcal{W}.\mathrm{H_j}.\mathcal{M}_{\mathrm{j,i}} \ .$$

The first term counts all words ending with $\mathrm{H_i}$; the second term enumerates the words in $\mathcal{W}.\mathrm{H_i}$ that contain at least one additional occurrence $\mathrm{H_j}$ from $\mathcal{H}$. Such words are not in $\mathcal{R}_i$. Notice that these occurrences are always valid.

In the renewal scheme, one must also subtract from $\mathcal{W}.\mathcal{H}_i$ the set of words with several occurrences from $\mathcal{H}$, only one being valid; this leads to:

$$\forall i : \mathcal{R}_i = \mathcal{W}.\mathcal{H}_i - \sum_j \mathcal{W}.\mathrm{H_j}.\mathcal{M}_{\mathrm{j,i}} - \sum_{\mathrm{j}} \mathcal{R}_{\mathrm{j}}(\mathcal{A}_{\mathrm{j,i}} - \{\epsilon\}) \ ;$$

notice that $\epsilon$ is in $\mathcal{A}_{j,i}$ iff $i = j$. Hence, this equation is equivalent to:

$$\forall i : \sum_j \mathcal{R}_j \mathcal{A}_{j,i} = \mathcal{W}.\mathcal{H}_i - \sum_j \mathcal{W}.\mathcal{H}_j.\mathcal{M}_{j,i} \ .$$

**Proposition 7** *In the overlapping scheme, the following language equations hold:*

$$\forall (i,j) : \sum_{k \geq 1} \mathcal{M}_{i,j}^{(k)} = \mathcal{W}\mathrm{H_j} + (\mathcal{A}_{\mathrm{i,j}} - \{\epsilon\}) \ . \tag{4}$$

*In the renewal scheme, the following language equations hold:*

$$\forall (i,j) : \sum_{k \geq 1} \sum_{l=1}^{q} \mathcal{M}_{i,l}^{(k)} \times \mathcal{A}_{l,j} = \mathcal{W}\mathrm{H_j} \ . \tag{5}$$

**PROOF.** Let $w$ be in $\mathcal{W}\mathrm{H_j}$ and let $k+1$ be the number of valid $\mathcal{H}$ occurrences in $\mathrm{H_i w}$ when $\mathrm{H_i}$ is valid. In both models, $k+1$ is greater than or equal to 1: if suffix $\mathrm{H_j}$ is not valid (renewal scheme), then it overlaps with a valid occurrence $\mathrm{H_l}$. Such an occurrence cannot overlap with valid occurrence $\mathrm{H_i}$, hence it is a factor of $w$ and $w$ rewrites $mt$ where prefix $m$ is in $\mathcal{M}_{i,l}^{(k)}$. In the overlapping scheme, suffix $\mathrm{H_j}$ is valid, hence $t$ is empty and $m$ is in $\mathcal{M}_{i,j}^{(k)}$. In the renewal scheme, when $t$ is not the empty string, it is a proper suffix of $\mathrm{H_j}$. Otherwise, by the reasoning above, it would contain an additional valid occurrence. Hence $t$ is in $\mathcal{A}_{l,j}$.

In examples above, basic languages were given explicitly. It is noteworthy that, in general, equations (2)-(5) do not provide explicit expressions for sets $\mathcal{R}_i, \mathcal{M}_{i,j}, \mathcal{U}_i$ e.g. cannot be inverted in general. Nevertheless, they appear adapted for enumeration purposes, that are developed in the next section.

## 4 Generating Functions

### *4.1 Definitions*

One combines the probability generating functions and the ordinary generating functions used in combinatorics for enumeration purposes. In this problem, the combinatorial data structures that appear are languages. Consider first the probability generating functions involved in this problem.

**Definition 8** *Given a set $\mathcal{H}$ of patterns $(\mathrm{H}_1, \ldots, \mathrm{H}_q)$ searched in a random text of size $n$, one denotes $N_{\mathrm{H}_i}$ the random variable that counts the number of* valid occurrences *of pattern $\mathrm{H}_i$ in a random text $t$. Conditioning by the size $n$ of the text yields random variables $N_{\mathrm{H}_i}(n)$.*

**Definition 9** *Given a set $\mathcal{H}$ of patterns $(\mathrm{H}_1, \ldots, \mathrm{H}_q)$ the probability generating function $P_n(u_1, \cdots, u_q)$ associated to the $q$-uple of random variables $(N_{\mathrm{H}_1}(n), \cdots, N_{\mathrm{H}_q(n)})$ is:*

$$P_n(u_1, \cdots, u_q) = \sum_{r_1, \cdots, r_q} \Pr(N_{\mathrm{H}_1}(n) = r_1, \cdots, N_{\mathrm{H}_q(n)} = r_q) u_1^{r_1} \cdots u_q^{r_q} \quad (6)$$

Exists a simple relation between the moments of these random variables and the derivatives of $P_n(u_1, \ldots, u_q)$ at $(u_1, \cdots, u_q) = (1, \cdots, 1)$. Namely [11]:

$$E(N_{\mathrm{H}_i}(n)) = \frac{\partial P_n}{\partial u_i}(1, \cdots, 1) \quad (7)$$

$$E(N_{\mathrm{H}_i}(n) N_{\mathrm{H}_j}(n)) = \frac{\partial^2 P_n}{\partial u_i \partial u_j}(1, \cdots, 1) \quad (8)$$

**Remark 10** *Un-conditioning allows easier computation through complex analysis. This leads to a combination with ordinary generating functions used in combinatorics [13].*

**Definition 11** *For any language $\mathcal{L}$ its* generating function *$L(z)$ is defined as*

$$L(z) = \sum_{w \in \mathcal{L}} P(w) z^{|w|} \quad (9)$$

*where $|w|$ is the length of $w$, with the usual convention that $P(\epsilon) = 1$.*

*Given a pattern* H, *its* H-conditional generating function *is defined as:*

$$L_{[\mathrm{H}]}(z) = \sum_{w \in \mathcal{L}} P(w|\mathrm{H}) z^{|w|} \ . \quad (10)$$

**Definition 12** *Given a set of patterns* $\mathcal{H} = (\mathrm{H_i})_{\mathrm{i=1\cdots q}}$, *the* multivariate generating function *for the number of occurrences is defined as:*

$$T(z, u_1, \ldots, u_q) = \sum_n z^n P_n(u_1, \cdots, u_q) \ . \tag{11}$$

**Notation:** We denote by $[z^n]f(z, u_1, \cdots, u_q)$ the coefficient of $z^n$ in the multivariate function $f$ and $[z^n]L(z)$ represents the coefficient of $z^n$ in the generating function $L(z)$.

In Definition 2, tail languages $\mathcal{U}_i$ and minimal languages $\mathcal{M}_{i,j}$ determine words that appear right of a given word $\mathrm{H_i}$. Hence, $\mathrm{H_i}$-conditional generating functions of $\mathcal{U}_i$ and $\mathcal{M}_{i,j}$ arise naturally.

**Definition 13** *One denotes* $M_{i,j}(z)$ *and* $U_{\mathrm{H_i}}(z)$ *the* $\mathrm{H_i}$*-conditional generating functions of languages* $\mathcal{M}_{i,j}$ *and* $\mathcal{U}_i$. *One defines:*

$$\mathbf{U}^t(z) = (..., U_{\mathrm{H_i}}(z), ...) \ ,$$
$$\mathbf{H}^t(z) = (P(\mathrm{H_1})\mathrm{z}^{\mathrm{m_1}}, \ldots, \mathrm{P(H_q)z}^{\mathrm{m_q}}) \ .$$

*Additionally,* $\mathbb{H}$ *is the* $q \times q$ *matrix with* $q$ *identical rows that are equal to* $\mathbf{H}^t(z)$. *Finally,* $\mathbb{M}(z)$ *is the* $q \times q$ *matrix which has* $M_{i,j}(z)$ *as its* $(i,j)$*-element, and the matrix associated to the minimal languages is:*

$$\mathbb{M}(z, u_1, \ldots, u_q) = (M_{i,j}(z)u_j) \ . \tag{12}$$

*One denotes* $\mathbb{M}(z) = \mathbb{M}(z, 1, \ldots, 1)$.

Initial languages appear as prefixes of the text sequences. Hence, the (unconditional) generating functions arise naturally.

**Definition 14** *One defines the row vector associated to initial languages:*

$$\mathbf{R}^t(z, u_1, \cdots, u_q) = (..., R_i(z)u_i, ...) \ . \tag{13}$$

*One denotes* $\mathbf{R}^t(z) = \mathbf{R}^t(z, 1, \cdots, 1)$ . *Finally,* $\mathbb{R}(z)$ *is the* $q \times q$ *matrix with* $q$ *rows identical to* $\mathbf{R}^t(z)$.

### *4.2 Basic generating functions*

Equations on basic languages will translate onto equations on their generating functions. Solving such equations in 4.3 will involve the generating functions

of some basic sets, that are derived in this section. The following notations appear useful:

**Definition 15** *Let $\mathbb{F}(z)$ be the $q \times q$ matrix defined by:*

$$\mathbb{F}(z)_{i,j} = \frac{1}{\pi_{\mathrm{H_j[1]}}}[(\mathbb{P}-\Pi)(\mathbb{I}-(\mathbb{P}-\Pi)z)^{-1}]_{\mathrm{H_i[m_i],H_j[1]}} \ ,$$

*where* $\mathrm{H_j[1]}$ *denotes the first character of* $\mathrm{H_j}$ *and* $\mathrm{H_i[m_i]}$ *denotes the last character of* $\mathrm{H_i}$*.*

It is noteworthy that $\mathbb{F}(z)$ is the zero matrix in the Bernoulli model.

**Proposition 16** *Let $\mathcal{W}$ denote the language of all words on a given alphabet $\mathcal{S}$. In the Bernoulli and Markov models, its generating function and its* H*-generating function satisfy, for any* H*:*

$$W(z) = W_{[\mathrm{H}]}(z) = \frac{1}{1-z} \ . \tag{14}$$

**PROOF.** From the definition, $W(z) = \sum_{w\in\mathcal{W}} P(w) z^{|w|} = \sum_{n=0}^{\infty} \sum_{|w|=n} P(w) z^n = \sum_{n=0}^{\infty} z^n = \frac{1}{1-z}$. The derivation of $W_{[\mathrm{H}]}(z)$ relies on the fact that for any probability matrix $\mathbb{P}$ and any character $i$, one has $\sum_{j\in\mathcal{S}} \mathbb{P}_{i,j} = 1$. One has:

$$W_{[\mathrm{H}]}(z) = \sum_{n=0}^{\infty} \sum_{|w|=n} P(w|\mathrm{H}) \mathrm{z^n} = 1 + \sum_{\mathrm{n=1}}^{\infty} \sum_{\mathrm{j=1}}^{\mathrm{V}} \mathbb{P}^{\mathrm{n}}_{\mathrm{H[m],j}} \mathrm{z^n} = 1 + \sum_{\mathrm{n=1}}^{\infty} \mathrm{z^n}$$

**Proposition 17** *The generating function of the set $\mathcal{W}.\mathrm{H_j}$ is:*

$$\frac{1}{1-z} P(\mathrm{H_j}) \mathrm{z}^{|\mathrm{H_j}|} \ .$$

*The* $\mathrm{H_i}$*-conditional generating function of the set $\mathcal{W}.\mathrm{H_j}$ is:*

$$(\frac{1}{1-z} + \mathbb{F}(z)_{i,j}) \times P(\mathrm{H_j}) \mathrm{z}^{|\mathrm{H_j}|} = (\frac{1}{1-\mathrm{z}} \mathbb{H}(\mathrm{z}) + \mathbb{F}(\mathrm{z})\mathbb{H}(\mathrm{z}))_{\mathrm{i,j}}$$

**PROOF.** The generating function for $\mathcal{W}\mathrm{H}$ is: $\sum_n z^n [\boldsymbol{\pi}\mathbb{P}^n]_{H[1]} \times \frac{P(\mathrm{H})}{\boldsymbol{\pi}_{\mathrm{H}[1]}} z^{|H|}$. Applying $n$ times the stationarity equation $\boldsymbol{\pi}\mathbb{P} = \boldsymbol{\pi}$ yields $[\boldsymbol{\pi}\mathbb{P}^n]_{H[1]} = \boldsymbol{\pi}_{H[1]}$ and the result follows. The $\mathrm{H_i}$-conditional generating function for $\mathcal{W}.\mathrm{H_j}$ is: $\sum_{n\geq 1} \mathbb{P}^n_{\mathrm{H_i[m_i],H_j[1]}} z^n \times \frac{P(\mathrm{H_j})}{\boldsymbol{\pi}_{\mathrm{H_j}[1]}} z^{|\mathrm{H_j}|-1}$. Rewriting $\mathbb{P}^n = \Pi^n + (\mathbb{P}-\Pi)^n$ yields the result.

By the methods given in [13], the translation of (1) into an equation on the multivariate generating function is "automatic" :

**Theorem 18** *Given a set $\mathcal{H}$ of patterns, the multivariate generating function $T(z, u_1, \ldots, u_q)$ satisfies the fundamental equation:*

$$T(z, u_1, \ldots, u_q) = \mathbf{R}^t(z, u_1, \cdots, u_q) \times (\mathbb{I} - \mathbb{M}(z, u_1, \ldots, u_q))^{-1} \times \mathbf{U}(z) \ . (15)$$

**PROOF.** Let us compute the contribution to the generating function $T(z, u_1, \ldots, u_q)$ of a word $t = w_{i_1} w_{i_2} \ldots w_{i_k} u$ in $\mathcal{L}_k$. The probability that $t$ occurs is: $P(w_{i_1})P(w_{i_2}|w_{i_1}) \cdots P(w_{i_k}|w_{i_{k-1}})P($
One observes that $P(w_{i_j}|w_{i_{j-1}}) = P(w_{i_j}|\mathrm{H_{i_{j-1}}})$ and $P(u|w_{i_k}) = P(u|\mathrm{H_{i_k}})$. Additionally, $w_{i_1}$ is in $\mathcal{R}_{i_1}$, $w_{i_2}$ is in $\mathcal{M}_{i_1,i_2}, \ldots,$ $w_{i_k}$ is in $\mathcal{M}_{i_{k-1},i_k}$. Now, $z^{|t|}$ rewrites $z^{|w_{i_1}|} \ldots z^{|w_{i_k}|} z^{|u|}$ and for a given subset $\{i_1, \ldots, i_k\}$, the associated monomial is $u_{i_1} u_{i_2} \cdots u_{i_k}$. Reordering yields $z^{|w_{i_1}|} P(w_{i_1}) u_{i_1} z^{|w_{i_2}|} P(w_{i_2}|\mathrm{H_{i_1}}) \mathrm{u_{i_2}} \cdots \mathrm{P(u|H_{i_k})z^{|u|}}$. Summation over possible decompositions rewrites:

$$\sum_{w_{i_1} \in \mathcal{R}_{i_1}} z^{|w_{i_1}|} P(w_{i_1}) u_{i_1} \sum_{w_{i_2} \in \mathcal{M}_{i_1,i_2}} z^{|w_{i_2}|} P(w_{i_2}|\mathrm{H_{i_1}}) \mathrm{u_{i_2}} \ldots \sum_{\mathrm{u} \in \mathcal{U}_{\mathrm{i_k}}} \mathrm{P(u|H_{i_k})z^{|u|}} \ ,$$

which is:

$$R_{i_1}(z) u_{i_1} M_{i_1,i_2}(z) u_{i_2} \ldots M_{i_{k-1},i_k}(z) u_{i_k} U_{i_k}(z) \ .$$

Summing over all subsets $\{i_1, \ldots, i_k\}$ yields: $\mathbf{R}^t(z, u_1, \cdots, u_q) \times \mathbb{M}(z, u_1, \ldots, u_q)^{k-1} \times \mathbf{U}(z)$ and summing over all $k$ gives the result.

We now state our main theorem for minimal languages. Notice that, in the renewal case, all minimal languages $\mathcal{M}_{i,j}$ are equal to the corresponding initial language $\mathcal{R}_j$.

**Theorem 19** *The generating function of the minimal languages satisfy the following matrical equations:*

*(a) Overlapping scheme:*

$$(\mathbb{I} - \mathbb{M}(z))^{-1} = \mathbb{A}(z) + (\frac{1}{1-z} + \mathbb{F}(z))\mathbb{H}(z) \ ; \qquad (16)$$

*(b) Renewal scheme:*

$$(\mathbb{I} - \mathbb{M}(z))^{-1} = \mathbb{I} + (\frac{1}{1-z} + \mathbb{F}(z))\mathbb{H}(z)\mathbb{A}(z)^{-1} \ . \qquad (17)$$

**PROOF.** One uses the $\mathrm{H_i}$-conditional generating function for $\mathcal{W}\mathrm{H_j}$ derived in Proposition (17). For any $(i,j)$, the generating function of the right-hand side of (4) in Proposition (7) is: $[\mathbb{A}(z) + (\frac{1}{1-z} + \mathbb{F}(z))\mathbb{H}(z)]_{i,j}$. One associates $\sum_k \mathbb{M}^k(z)_{i,j}$ to the left-hand side. Hence, we get matrical equation (16). Equation (17) follows similarly from (5).

**Theorem 20** *The generating functions of the initial languages satisfy the following matrical equation:*

$$\mathbf{R}^t(z) = \frac{1}{1-z}\mathbf{H}^t(z) \times [\mathbb{A}(z) + (\frac{1}{1-z} + \mathbb{F}(z))\mathbb{H}(z)]^{-1} \quad . \tag{18}$$

**PROOF.** This follows directly from equations above.

Finally, tail languages satisfy:

**Proposition 21** *In Bernoulli and Markov models, for overlapping or non-overlapping occurrences, the generating functions of the tail languages satisfy the matrical equation:*

$$\mathbf{U}(z) = \begin{pmatrix} \cdots \\ U_{\mathrm{H_i}}(z) \\ \cdots \end{pmatrix} = (\mathbb{I} - \mathbb{M}(z)) \times \frac{1}{1-z} \times \mathbf{1_q} \quad . \tag{19}$$

**PROOF.** It follows from Equation (14) in Proposition (16) that:

$$U_{\mathrm{H_i}}(z) = W_{[\mathrm{H_i}]}(z) - \sum_j M_{i,j}(z) W_{[\mathrm{H_j}]}(z) = (1 - \sum_j M_{i,j}(z))\frac{1}{1-z} \quad .$$

As $\sum_j M_{i,j}(z)$ is the $i$-th row of $\mathbb{M}(z) \times \mathbf{1_q}$, the result follows.

Although (19) does not depend on the model, observe that $(U_{\mathrm{H_i}}(z))$ depend on it through $\mathbb{M}(z)$. It will appear below that (19) is enough for the main purpose of this paper and there is no need for an explicit expression of $(U_{\mathrm{H_i}}(z))$. Nevertheless, observe that plugging (16) or (17) into (19) yields a set of equations for $U_{\mathrm{H_i}}(z)$ for each model.

## 5 Mean, Variances and Covariances

A challenging point is the computation of the mean, variance and covariances. Symbolic computation appears here a very powerful tool that allows a computation in the finite range at the same computational effort as an asymptotic computation.

More precisely, Equations (7)-(8) rewrite:

**Lemma 22**

$$E(N_{\mathrm{H_i}}(n)) = [z^n]\frac{\partial T}{\partial u_i}(z,1,\cdots,1) \tag{20}$$

$$Cov(N_{\mathrm{H_i}}(n), N_{\mathrm{H_j}}(n)) = [z^n](\frac{\partial^2 T}{\partial u_i \partial u_j} - \frac{\partial T}{\partial u_i}\frac{\partial T}{\partial u_j})(z,1,\cdots,1) \tag{21}$$

$$Var(N_{\mathrm{H_i}}(n)) = [z^n](\frac{\partial^2 T}{\partial u_i^2} + \frac{\partial T}{\partial u_i} - (\frac{\partial T}{\partial u_i})^2)(z,1,\cdots,1) \tag{22}$$

**PROOF.** A term by term derivation of (11) yields:

$$\frac{\partial T}{\partial u_i}(z,u_1,\ldots,u_q) = \sum_n z^n \frac{\partial P_n(u_1,\ldots,u_q)}{\partial u_i}$$

It follows from (7) that:

$$\frac{\partial T}{\partial u_i}(z,1,\cdots,1) = \sum_n z^n \frac{\partial P_n}{\partial u_i}(z,1,\cdots,1) = \sum_n z^n E(N_{\mathrm{H_i}}(n))$$

and this is (20). Results on variances and covariances follow similarly from (8).

It is noteworthy that Lemma 22 allows to avoid a formal inversion of Equation (15). Only the derivatives at 1 are needed to compute the moments and it appears that simplifications of the derivatives at $(u_1,\cdots,u_q) = (1,\cdots,1)$ are expected. This will be developed in the next subsections.

### *5.1 Formal expressions*

The aim of this subsection is the derivation of Theorem 23 below. It provides formal expressions for the partial derivatives of $T$ that appear in Lemma 22, as functions of the generating functions of basic languages.

**Theorem 23** *The row vector* $(\ldots, \frac{\partial T}{\partial u_i}(z,1,\ldots,1),\ldots)$ *of partial derivatives is:*

$$\frac{1}{1-z}\mathbf{R}^t(z) \times (\mathbb{I} - \mathbb{M}(z))^{-1} \tag{23}$$

*The matrix of second derivatives satisfies:*

$$(\frac{\partial^2 T}{\partial u_i \partial u_j}(z,1,\ldots,1)) = \frac{1}{1-z}(\mathbb{D}(z) + \mathbb{D}(z)^t) \ . \tag{24}$$

*where*

$$\mathbb{D} = (\mathbb{R}(z)((\mathbb{I} - \mathbb{M}(z))^{-1})).((\mathbb{I} - \mathbb{M}(z))^{-1\,t} - \mathbb{I}) \ .$$

**Remark 24** *In the Bernoulli model and the renewal case, (24) reduces to*

$$\frac{\partial^2 T}{\partial u_i \partial u_j}(z,1,\ldots,1) = \frac{1}{1-z}\frac{R_i(z)R_j(z)}{(1-R_{\mathcal{H}}(z))^2}$$

*which can be obtained by a direct derivation.*

The proof relies on the following lemma:

**Lemma 25** *For any* $i$*, the following result holds:*

$$\frac{\partial(\mathbb{I} - \mathbb{M}(z,u_1,\ldots,u_q))^{-1}}{\partial u_i} = (\mathbb{I} - \mathbb{M}(z,u_1,\ldots,u_q))^{-1} \times \mathbb{M}(z)_i \times (\mathbb{I} - \mathbb{M}(z,u_1,\ldots,u_q))^{-1} \ .$$

**PROOF.** Remark first that $\frac{\partial \mathbb{M}(z,u_1,\ldots,u_q)}{\partial u_i}$ is the matrix $\mathbb{M}(z)_i$. Now, one derives $\mathbb{M}^k \sum_{\ell=0}^{k-1} \mathbb{M}^\ell \dot{\mathbb{M}} \dot{\mathbb{M}}^{k-(\ell+1)}$ as the sum:

$$\frac{\partial \mathbb{M}(z,u_1,\ldots,u_q)^k}{\partial u_i} = \sum_{\ell=0}^{k-1} \mathbb{M}(z,u_1,\ldots,u_q)^\ell \mathbb{M}(z)_i \mathbb{M}(z,u_1,\ldots,u_q)^{k-(\ell+1)} \ .$$

Then, a term by term derivation of $(\mathbb{I} - \mathbb{M}(z,u_1,\ldots,u_q))^{-1} = \sum_k \mathbb{M}(z,u_1,\ldots,u_q))^k$ yields

$$\sum_\ell \sum_m \mathbb{M}(z,u_1,\ldots,u_q)^\ell \mathbb{M}(z)_i \mathbb{M}(z,u_1,\ldots,u_q)^m$$

Grouping yields finally $(\mathbb{I} - \mathbb{M}(z,u_1,\ldots,u_q))^{-1}\mathbb{M}(z)_i(\mathbb{I} - \mathbb{M}(z,u_1,\ldots,u_q))^{-1}$.

One can now proceed with the proof of the theorem:

**PROOF (Theorem)** One uses Lemma 25 to derive $\frac{\partial T}{\partial u_i}(z, u_1, \ldots, u_q)$ in (15). Additionally, $\frac{\partial \mathbf{R}^t(z,u_1,\cdots,u_q)}{\partial u_i}(z, 1, \ldots, 1)$ is a row vector where the $i$-th term is $R_i(z)$ and other terms are 0. It can be rewritten $\mathbf{R}^t(z, 1, \cdots, 1) \times \mathbb{I}_i$. Hence, $\frac{\partial T}{\partial u_i}(z, u_1, \ldots, u_q)$ is:

$$\begin{aligned}&[\ \mathbf{R}^t(z) \times \mathbb{I}_i + \mathbf{R}^t(z, u_1, \cdots, u_q)(\mathbb{I} - \mathbb{M}(z, u_1, \ldots, u_q))^{-1}\mathbb{M}(z)_i] \\ &\times (\mathbb{I} - \mathbb{M}(z, u_1, \ldots, u_q))^{-1}\mathbf{U}(z) \end{aligned} \tag{25}$$

One rewrites: $(\mathbb{I}-\mathbb{M}(z, u_1, \ldots, u_q))^{-1}\mathbb{M}(z)_i = \sum_{k\geq 0} \mathbb{M}(z, u_1, \ldots, u_q)^k \mathbb{M}(z, u_1, \ldots, u_q)_i$. When $(z, u_1, \ldots, u_q) = (1, \ldots, 1)$, this is $\sum_{k\geq 0}[\mathbb{M}(z)^{k+1}]_i = (\mathbb{I} - \mathbb{M}(z))_i^{-1} - \mathbb{I}_i$. Now (19) implies that:

$$(\mathbb{I} - \mathbb{M}(z, 1, \ldots, 1))^{-1} \times \mathbf{U}(z) = (\mathbb{I} - \mathbb{M}(z))^{-1}\mathbf{U}(z) = \frac{1}{1-z}\mathbf{1_q}\ .$$

It follows that: $\frac{\partial T}{\partial u_i}(z, 1, \ldots, 1) = \mathbf{R}^t(z) \times (\mathbb{I} - \mathbb{M}(z))_i^{-1} \times \frac{1}{1-z}\mathbf{1_q}$ and the result follows for the mean.

Variances and covariances come out the same**.** Derivation with respect to $u_j$ of the second factor of (25) yields:

$$(\mathbb{I} - \mathbb{M}(z, u_1, \ldots, u_q))^{-1}\mathbb{M}(z)_j \times (\mathbb{I} - \mathbb{M}(z, u_1, \ldots, u_q))^{-1}\mathbf{U}(z)\ .$$

A derivation with respect to $u_j$ of the first factor yields:

$$\begin{aligned}&[\mathbf{R}^t(z, u_1, \cdots, u_q)(\mathbb{I} - \mathbb{M}(z, u_1, \ldots, u_q))^{-1}\mathbb{M}(z)_j + \mathbf{R}^t(z) \times \mathbb{I}_j] \\ &\times (\mathbb{I} - \mathbb{M}(z, u_1, \ldots, u_q))^{-1}\mathbb{M}(z)_i\end{aligned}$$

Term $(\mathbb{I} - \mathbb{M}(z, u_1, \ldots, u_q))^{-1}\mathbf{U}(z)$ factorizes and simplifies into $\frac{1}{1-z}\mathbf{1_q}$ when $(z, u_1, \ldots, u_q) = (z, 1, \ldots, 1)$. When $(u_1, \cdots, u_q) = (1, \cdots, 1)$, $\mathbf{R}^t(z) \times \mathbb{I}_i + \mathbf{R}^t(z, u_1, \cdots, u_q) \times (\mathbb{I} - \mathbb{M}(z, 1, \ldots, 1))^{-1}\mathbb{M}_i$ simplifies into $\mathbf{R}^t(z) \times (\mathbb{I} - \mathbb{M}(z))_i^{-1}$ and this provides $\mathbf{R}^t(z) \times (\mathbb{I} - \mathbb{M}(z))_i^{-1} \times (\mathbb{I} - \mathbb{M}(z))^{-1}\mathbb{M}_j$. Observe that, for any matrix $\mathbb{M}$, $\mathbb{M}_i \times \mathbb{I}_j = 0$ when $i \neq j$. Then, this term reduces to $\mathbf{R}^t(z) \times (\mathbb{I} - \mathbb{M}(z))_i^{-1} \times (\mathbb{I} - \mathbb{M}(z))_j^{-1}$. Now, the contribution of $(\mathbb{I} - \mathbb{M}(z))_j^{-1}$ in this product is a multiplication by the $(i, j)$-th element of $(\mathbb{I} - \mathbb{M}(z))^{-1}$; $\mathbf{R}^t(z) \times (\mathbb{I} - \mathbb{M}(z))_i^{-1}$ is the $i$-th element of row vector $\mathbf{R}^t(z) \times (\mathbb{I} - \mathbb{M}(z))^{-1}$ which also is the $(j, i)$ element of matrix $||\mathbb{R}(z)(\mathbb{I}-\mathbb{M}(z))^{-1}||$ and we get $\mathbb{D}(z)$. One must now subtract, when $i = j$, the contribution $\mathbf{R}^t(z) \times (\mathbb{I} - \mathbb{M}(z))_{i,i}^{-1}$. Summing over all $i$, this yields $\mathbf{R}^t(z) \times (\mathbb{I} - \mathbb{M}(z))^{-1}.\mathbb{I}$.

The remaining terms provide the symmetric term.

*5.2 Practical computation*

At this stage, one plugs the equations derived in Section 4 for each specific counting scheme into equations (23) and (24).

**Theorem 26** *Let $\mathcal{H}$ be a set of patterns* $\mathrm{H_i}$ *of sizes $m_i$. In the Markov and Bernoulli case, in the renewal model, the row vector of expectations is:*

$$n\mathbf{H}(1)^t\mathbb{A}(1)^{-1} + [\mathbf{H}(1)^t\mathbb{A}(1)^{-1} + \mathbf{H}(1)^t\mathbb{A}(1)^{-1}\mathbb{A}'(1)\mathbb{A}(1)^{-1} - \mathbf{H}'(1)^t\mathbb{A}(1)^{-1}] \quad .(26)$$

*When $\mathcal{H}$ reduces to a single pattern* H, *this leads to the equation:*

$$E(N_H(n)) = \frac{P(\mathrm{H})}{\mathrm{A_H}(1)}[n - m + 1 + \frac{A'_{\mathrm{H}}(1)}{A_{\mathrm{H}}(1)}] \ . \qquad (27)$$

The following result, that can be obtained in a direct manner, is well-known [33]:

**Expectation in the overlapping case:** The expected value of the number of occurrences of a given pattern $\mathrm{H_i}$ is:

$$E(N_{\mathrm{H_i}}(n))) = (n - m_i + 1)P(\mathrm{H_i}) \ . \qquad (28)$$

**Remark 27** *The linear term in the renewal model coincides with the one given in [6]. Also, when a word* $\mathrm{H_i}$ *is not self-overlapping, then its autocorrelation polynomial $A_{\mathrm{H_i}}(z) = 1$ and this result is consistent with the overlapping model result.*

**PROOF.** It relies on the following lemma:

**Lemma 28** *The row vector of partial derivatives is equal to:*

*(i)* $\frac{1}{(1-z)^2}\mathbf{H}^t(z)$ *in the overlapping scheme;*
*(ii)* $\frac{1}{(1-z)^2}\mathbf{H}^t(z)\mathbb{A}(z)^{-1}$ *19 in the renewal scheme.*

**PROOF.** One plugs (16), (17) and (18) into (23).

In the renewal case, a Taylor expansion at $z = 1$ yields:

$$\mathbf{H}^t(z)\mathbb{A}(z)^{-1} = \mathbf{H}^t(1)\mathbb{A}(1)^{-1} + (1-z)[\mathbf{H}^t(1)\mathbb{A}^{-1}(1)\mathbb{A}'(1)\mathbb{A}^{-1}(1) - \mathbf{H}'(1)^t\mathbb{A}(1)^{-1}] \ .$$

Then, one applies the classical formula that generalises the famous binomial equation:

$$[z^n](1-z)^{-p} = \frac{\Gamma(n+p)}{\Gamma(p)\Gamma(n+1)} \tag{29}$$

where $[z^n]$ means the $z^n$ coefficient and $\Gamma$ represents the *Gamma* function that satisfies $\Gamma(n+1) = n!$ when $n$ is an integer. This yields (26).

It is noteworthy that $\mathbb{A}(z), \mathbf{H}^t(z)$ and derivatives at $z = 1$ arise at this step, crucial to avoid the inversion of polynomial matrix $\mathbb{A}(z)$, hence to lower the computational complexity.

Method applies for the overlapping model. Then, the $i$-th element in the row vector of expectations is: $\frac{1}{(1-z)^2}P(\mathrm{H_i})\mathrm{z}^{\mathrm{m_i}}$. Applying (29), one gets (28).

**Remark 29** *Intuitively, when a text is long enough, the probability p to find a valid occurrence at a given position does not depend of the position. Hence, the linearity constant is p, that depends on the counting model. The constant term arises from end effects. First, in both counting models, the pattern cannot appear in the* $(m-1)$ *last positions, and one subtract* $(m-1)p$*. A second end effect, rather subtle, also appears in the renewal model: the dependance to the past. Namely, the validity of an occurrence at position i depends on the chain of overlapping occurrences ending at position i, if any. Length* $\ell$ *of such a chain is upper bounded by i. Term* $\mathbf{H}(1)^t\mathbb{A}(1)^{-1}\mathbb{A}'(1)\mathbb{A}(1)^{-1}$ *accounts for this truncation. This is easily checked when* $\mathcal{H}$ *reduces to a singleton, as one subtracts:*

$$P(\mathrm{H}) \sum_{\mathrm{i=1}}^{\mathrm{n-m+1}} \sum_{\ell \geq \mathrm{i}} [\mathrm{z}^\ell] \cdot \frac{1}{\mathrm{A_H(z)}}$$

*which tends to* $\frac{P(\mathrm{H})\mathrm{A'_H}(1)}{\mathrm{A^2_H}(1)}$ *when n tends to infinity. Approximation order is exponentially small. This follows rigorously from the analytic approach, or, more intuitively, from the fact that the probability of an overlapping chain of length* $\ell$ *decreases exponentially with* $\ell$*.*

**Theorem 30** *Let* $\mathcal{H}$ *be a set of patterns* $\mathrm{H_i}$ *of sizes* $m_i$*. Let* $\mathbb{B}(z), \mathbb{C}(z)$*,* $\mathbb{E}(z)$ *and* $\mathbb{L}(z)$ *be the matrices:*

*(i) Overlapping case:*

$$\mathbb{L}(z) = \mathbb{H}(z)$$
$$\mathbb{C}(z) = \mathbb{H}(z).(\mathbb{A}(z) - \mathbb{I}) \ ,$$

*(ii) Renewal case:*

$$\mathbb{L}(z) = \mathbb{H}(z)\mathbb{A}(z)^{-1}$$
$$\mathbb{C}(z) = 0 \ ,$$

*and, in both cases:*

$$\mathbb{B}(z) = \mathbb{L}(z).\mathbb{L}(z)^t$$
$$\mathbb{E}(z) = \mathbb{L}(z).(\mathbb{F}(z)\mathbb{L}(z))^t \ .$$

*The variance-covariance matrix is equal to:*

$$n[\mathbf{X}_1 + \mathbf{X}_2] + [\mathbf{Y}_1 + \mathbf{Y}_2] \ , \tag{30}$$

*where:*

$$\mathbf{X}_1 = \mathbb{B}(1) - \mathbb{B}'(1) + \mathbb{C}(1) + Diag(\mathbb{L}(1))$$
$$\mathbf{Y}_1 = \mathbf{X}_1 + \mathbb{B}\text{“}(1) - \mathbb{L}'(1).\mathbb{L}'(1)^t - \mathbb{C}'(1) - Diag(\mathbb{L}'(1))$$
$$\mathbf{X}_2 = \mathbb{E}(1) + \mathbb{E}(1)^t$$
$$\mathbf{Y}_2 = \mathbf{X}_2 - (\mathbb{E}'(1) + \mathbb{E}'(1)^t) \ ,$$

*In both cases,* $\mathbf{X}_2$ *and* $\mathbf{Y}_2$ *reduce to* 0 *in the Bernoulli model.*

**PROOF.** In the overlapping model, it follows from (16) and (18) that $\mathbb{D}(z) = \frac{1}{1-z}\mathbb{H}(z).(\mathbb{A}(z) - \mathbb{I} + (\frac{1}{1-z} + \mathbb{F}(z))\mathbb{H}(z))^t$. This rewrites $\frac{1}{(1-z)^2}\mathbb{H}(z).\mathbb{H}(z)^t + \frac{1}{1-z}\mathbb{H}(z).(\mathbb{A}(z) - \mathbb{I} + \mathbb{F}(z)\mathbb{H}(z))^t$. In the renewal case, it follows from (17) and (18) that

$$\mathbb{D}(z) = \frac{1}{(1-z)^2}\mathbb{H}(z)\mathbb{A}(z)^{-1}.(\mathbb{H}(z)\mathbb{A}(z)^{-1})^t + \frac{1}{1-z}\mathbb{H}(z)\mathbb{A}(z)^{-1}.(\mathbb{F}(z)\mathbb{H}(z)\mathbb{A}(z)^{-1})^t \ .$$

In both cases, $\mathbb{D}(z)$ rewrites $\frac{1}{(1-z)^3}\mathbb{B}(z) + \frac{1}{(1-z)^2}\mathbb{C}(z) + \frac{1}{(1-z)^2}\mathbb{E}(z)$. One uses again (29), and a Taylor expansion at point $z = 1$ yields the asymptotic expansion of second partial derivatives. One must subtract the product of partial derivatives, e.g. :

$$((n+1)\mathbb{L}(1) - \mathbb{L}'(1)).((n+1)\mathbb{L}(1) - \mathbb{L}'(1))^t$$

A symbolic computation system such as Maple can output these formulae for the variance-covariance matrix, providing the linear term and the constant term with the same computational effort.

**Remark 31** *First results on variance in the overlapping scheme can be found in [25]. The problem was first addressed globally in the unpublished thesis [22] whose results are summarised in [33]. The matrix* $\mathbf{X}_1$ *was derived in [22] and, in a slightly more general case, in [2]. Additionally, the author of [22] expressed the Markovian term* $\mathbf{X}_2$ *(that is* 0 *in the Bernoulli case) as infinite sums. Expression in the theorem above provides the* simplification *of these infinite sums as a closed computable formula, as well as the constant terms. In [28], we addressed the case where* $\mathcal{H}$ *reduces to a single pattern, e.g. we computed this constant term in the variance in the Markov case. The key to the simplification over [22] that occurs is the quite general matrix equation* $(\mathbb{I}-\mathbb{M})^{-1}=\sum_{r\geq 0}\mathbb{M}^k$ *that transforms the computation of an infinite sum into the (less costly) inversion of a matrix. More precisely, the power of this approach is deeply related to the fact that "all" occurrences problems can be modellized by a finite state automaton, which in turn "creates" the simplification.*

**Remark 32** *Matrix* $\mathbf{X}_1$ *above was first derived for the renewal case in the Bernoulli model in [32].*

### *5.3 Computational complexity*

In all approaches, it appears that the computation of the mean and the variance imply, at some step, the *inversion of a linear system*. Depending on the approach, the size (or the structure) of the pattern set and the probabilistic model imply limits to the tractability of formulae. We discuss below the *minimal size* of the linear system involved.

It was proved above that all results depend on the overlapping of patterns in the set $\mathcal{H}$. A fundamental advantage of Theorem 30 is to provide formulae that are *computable*. More precisely, $\mathbb{A}(1)$ is a $q\times q$ matrix of real numbers (often rational in practice), while $\mathbb{A}(z)=\mathbb{D}(z)$ in [32] is a matrix of polynomials. Hence, the inversion of $\mathbb{A}(z)$ is costly and induces numerical instability while $\mathbb{A}(1)$ is still invertible in practice. Second, this formula shows that it is enough to compute $\mathbb{A}(1)^{-1},\mathbf{H},\mathbb{A}'(1)$ and $\mathbb{A}"(1)$ to derive the linar term and the constant term. Nevertheless, if set $\mathcal{H}$ is large, the computation of an autocorrelation matrix of size $|\mathcal{H}|\times|\mathcal{H}|$ [2,32,27] by brute force may lead to untractable formulae.

One first observe that not all patterns overlap with all other patterns. I.e. the matrix is sparse. In the case where the counting of each pattern separately is actually necessary, one may rely on this to derive the computation efficiently.

It appears more efficient, in the case where one counts all possible occurrences within the set $\mathcal{H}$ simultaneously, to *aggregate the states*:

**Definition 33** *Let $\mathcal{H}$ be a set of patterns. Let $(\mathcal{H}_i)_{1\leq i\leq q}$ be a* finite *partition satisfying the following property:*

$$\forall (i,j), \exists \mathcal{A}_{i,j} : \forall (\mathrm{H_i}, \mathrm{H_j}) \in \mathcal{H}_\mathrm{i} \times \mathcal{H}_\mathrm{j} : \textit{the correlation set of } \mathrm{H_i} \textit{ and } \mathrm{H_j} \textit{ is } \mathcal{A}_{i,j} \quad .(31)$$

**Fact 1:** $\mathcal{H}$ satisfies Property (31) iff:

$$\exists c_{i,j} : c_{i,j} \text{ is the maximal suffix of } \mathrm{H_i} \text{ that is a prefix of } \mathrm{H_j} \quad .$$

**Fact 2:** Given a set $\mathcal{H}$ and a partitioning satisfying the property above, the mean and variance of the expected number of $\mathcal{H}$-occurrences is linear and the linearity constants derived in Theorems 23, 26 and 30 hold.

The key observation is that any pattern $\mathrm{H_j}$ in $\mathcal{H}_j$ rewrites $c_{i,j}s_{i,j}$. Hence, the computation of $A_{i,j}(z)$ reduces to a summation: ($\sum P(s_{i,j})z^{|s_{i,j}|}$ in the Bernoulli case). This reduces the complexity of the derivation of the expected number of occurrences, that depend on the *size of the partition*. This size is briefly discussed on some examples below.

## 6 Numerical Evaluation

We now provide some numerical evaluations in the Bernoulli and Markov model, in the overlapping and renewal case. In order to make a comparison, we chose one example from [32,6] in the renewal case and Bernoulli model.

Let $\mathcal{H} = \{TTA, TAT, AA\}$. Then, in the overlapping case, the vector of expectations is:

$$\begin{array}{l}[(p_A - 2\,{p_A}^2 + {p_A}^3)\,n - 2\,p_A + 4\,{p_A}^2 - 2\,{p_A}^3,\\ (p_A - 2\,{p_A}^2 + {p_A}^3)\,n - 2\,p_A + 4\,{p_A}^2 - 2\,{p_A}^3,\\ {p_A}^2\,n - {p_A}^2]\end{array}$$

When $(p_A, p_T) = (0.5, 0.5)$, we get:

$$\begin{bmatrix} .125n - .250 \\ .125n - .250 \\ .25n - .25 \end{bmatrix}$$

When $(p_A, p_T) = (0.1, 0.9)$, we get:

$$\begin{bmatrix} .081n - .162 \\ .081n - .162 \\ .01n - .01 \end{bmatrix}$$

In the renewal case, we get the vector of expectations:

$$\left[ -\frac{(-1+p_A)^2\, p_A\, n}{{p_A}^3 - {p_A}^2 - 1} - \frac{(-1+p_A)^2\, p_A\, (p_A - 2\,{p_A}^2 + {p_A}^3 + 2)}{({p_A}^3 - {p_A}^2 - 1)^2}, \right.$$
$$-\frac{(-1+p_A)^2\, ppa\, ({p_A}^4 - {p_A}^3 - p_A)\, n}{({p_A}^3 - {p_A}^2 - 1)^2} - \frac{(-1+p_A)^2\, p_A\, (-1+3\,p_A)}{({p_A}^3 - {p_A}^2 - 1)^2},$$
$$\frac{({p_A}^7 - 3\,{p_A}^5 - {p_A}^4 + {p_A}^3 + 2\,{p_A}^2 + 2\,p_A)\, {p_A}^2\, n}{({p_A}^3 - {p_A}^2 - 1)^2\, (1+p_A)^2}$$
$$\left. + \frac{(-2\,{p_A}^5 + 4\,{p_A}^4 + 3\,{p_A}^3 - 5\,{p_A}^2 - 3\,p_A + 2)\, {p_A}^2}{({p_A}^3 - {p_A}^2 - 1)^2\, (1+p_A)^2} \right]$$

The numerical values are, when $(p_A, p_T) = (0.5, 0.5)$:

$$\begin{bmatrix} .1111111111n - .2098765432 \\ .05555555556n - .04938271605 \\ .1296296296n - .01646090535 \end{bmatrix}$$

The linear term coincide with the result derived in [32].

## 7 Miscelleanous Problems and Applications

### *7.1 Probabilities in the finite range*

It is also of interest to compute the distribution of the word count and the renewal count in the finite range. Determining words with unexpected frequencies implies the (fast) computation of the probability to find $r$ occurrences of a given word H in a text of size $n$, for *finite* $n$. It follows from Equation (15) and from language equations that this probability satisfies a *linear recurrence equation* of degree $rm$. More precisely, let us introduce $B^{(r)}(z)$ the generating

function of sequences that contain at least $r$ occurrences of a given word H. With the notations of Section 4, one gets, for any given $r$:

$$B^{(r)}(z) = \sum_{n\geq 0} P(N_{\mathrm{H}}(n) \geq r) z^n = \sum_{p\geq r} [u^p] T(z,u)$$
$$= \sum_{p\geq r-1} R(z) \cdot M_{\mathrm{H}}^p \cdot U_{\mathrm{H}}(z) = R(z) \cdot M_{\mathrm{H}}^{r-1} (1 - M_{\mathrm{H}}(z))^{-1} \cdot U_{\mathrm{H}}(z) \ .$$

Applying (19) yields the generating function: $\frac{1}{1-z} R(z) M_{\mathrm{H}}^{r-1}$. One now plugs (18) and, in the overlapping case, (16). Hence:

$$B^{(r)}(z) = \frac{P(\mathrm{H}) \mathrm{z}^{|\mathrm{H}|}}{1 - \mathrm{z}} \frac{[D_{\mathrm{H}}(z) - (1-z)]^{r-1}}{D_{\mathrm{H}}(z)^r} \tag{32}$$

with $D_{\mathrm{H}}(z) = (1-z) A_{\mathrm{H}}(z) + P(\mathrm{H}) \mathrm{z}^{|\mathrm{H}|} + (1 - \mathrm{z}) \mathrm{P}(\mathrm{H}) \mathrm{z}^{|\mathrm{H}|} \mathrm{F}(\mathrm{z})$. In the Bernoulli model, $D_{\mathrm{H}}(z)$ is a polynomial; hence, $[z^n] B^{(r)}(z) D_{\mathrm{H}}(z)^r (1-z)$ is 0. One rewrites $D_{\mathrm{H}}(z)^r (z)(1-z) = q_p z^p + q_{p-1} z^{p-1} + \cdots + q_0$. Then, $r_n = P(N_{\mathrm{H}}(n) \geq r)$ satisfies linear equation:

$$r_n q_0 + r_{n-1} q_1 + \cdots + r_{n-p} q_p = 0 \ .$$

This ensures a numerical computation that is stable and fast, e.g. $O(\log(n))$. Moreover, such an equation can be automatically written and solved by the software COMBSTRUCT developed by B. Salvy. Renewal model is treated the same. Markov model is trickier: additional term $(1-z) P(\mathrm{H}) \mathrm{z}^{|\mathrm{H}|} \mathrm{F}(\mathrm{z})$ introduces a correcting term that decreases exponentially. Implementation is currently done.

Parameters of interest for $r$-scans follow from a simplification and a differentiation of (15). For example, the probability of a first occurrence at position $\ell$ is $[z^\ell] R(z)$. Hence, average waiting time for the first occurrence is $R'(1)$, for the various probabilistic and counting models. Also, the method adapts to a modification of the constraints assumed: it only implies a modification of the equations on the languages defined in Definition 2.

**Other parameters of interest** to biologists can be studied through this approach. One can cite the search for Dos-DNA. The formulae above have been used by E. Coward [10]. A possible application is the use of the results above on covariances for the search of contrast words. To illustrate this suggestion, let us cite one application in [10]. One scan on a genome has shown that $CCG, CGA, GAC$ and $ACC$ were overrepresented. A high covariance was suggesting they were part of a bigger pattern. A more careful study has actually shown that $CCGA$ was appearing in tandem repeats.

An important application is the distribution of regular expressions. For example, homologous genes may be characterized in a database by a common motif, a profile, expressed as a regular expression. Such a profile has a very small probability to occur in a random gene but appears in all homologous genes of the family.

Regular expressions -that may represent infinite sets- are recognized by a finite automaton. This guarantees that the set $\mathcal{H}$ of all instantiations admits a partition satisfying Property (31), and all results derived above apply. In a recent work [24], adaptation of algorithms searching for regular expressions allows for the computation of mean and variance for any *given* regular expression. Nevertheless, closed formulae are not attainable and the size of the linear system to be inversed for a given regular expression is the size of the corresponding *searching automaton*. It is worth pointing out here that this size is always bigger than the size of a (minimal) partition satisfying Property (31). Let us illustrate the complexity improvements on one example.

Let $\mathcal{H}$ be the set of patterns that instantiate PROSITE expression PS00844:

$$[LIVM]x(3)[GA]x[GSAIV]R[LIVCA]D[LIVMF](2)x(7,9)[LI]xE[LIVA]N[STP]xP[GA]$$

Here [ ] stands for a choice and ( ) for a length. E.g. either one of the four characters $L, I, V, M$ can occur at position 1 . while the three next positions are unspecified; also, $x(7,9)$ means that 7 to 9 unspecified characters occur after position 13. The cardinality of this set is about $1.9 \times 10^{26}$. Inversion of the correlation matrix by brute force is intractable. Searching automaton size has, in the Bernoulli case, 946 states [24]. Nevertheless, set $\mathcal{H}$ can be partitioned into 5 states in Bernoulli and Markov models [27]. This rather surprising fact can be explained by the fact that, despite the number of unspecified choices, only a few overlaps are allowed. One expects this to be rather general on most PROSITE profiles.

Remark that in the very specific case where $\mathcal{H}$ is a set of non-overlapping patterns, overlapping matrix is diagonal. Equivalently, a 1-partition satisfies Property (31). Hence, all results in [28] steadily apply.

### *7.3 Approximate matching*

It appears of particular interest to be able to evaluate the expected number of *approximate* occurrences of a given word [19] in order to test the significance

of repeated approximate patterns. Among possible applications, let us cite the study of periodical patterns in sequences of promoters or to the search of regulatory sites. Let us illustrate our approach by a specific example. Let $H_0$ be the pattern *abacaba* and $\mathcal{H}$ the patterns which are within distance 1 according to Hamming distance (at most one substitution is allowed). It appears that $|\mathcal{H}| = 7 \times V$, where $V$ is the size of the alphabet $\mathcal{S}$. Nevertheless, $\mathcal{H}$ can be partitioned into 4 sets, in the Bernoulli case ($(4 + (q - 1))$ in the Markovian case) [27]. It is still an open problem to design an efficient algorithm to build the overlapping automaton in that case.

## 8 Conclusion

The problem of the counting of one or several patterns under various constraints was addressed here. A general scheme was provided allowing the derivation of exact formulae for the moments for Bernoulli and Markov model, that are linear functions of the size of the text. This work extends many previous results, and occasionnally simplifies them. Moreover, such computations can rely on symbolic computation systems with a low computational cost. This opens the way to practical uses of the formulae. First promising attempts can be found in [10].

## Acknowledgements

I am very grateful to A. Konopka and V. Makeev for their remarks and comments. It is also my pleasure to acknowledge all the valuable comments by the anonymous referees.

## References

[1] G.I. Bell and D.C. Torney. Repetitive DNA Sequences:Some Considerations for Simple Sequence Repeats. *Computers Chem.*, 17(2):185–190, 1993.

[2] Edward A. Bender and Fred Kochman. The Distribution of Subwords Counts is Usually Normal. *European Journal of Combinatorics*, 14:265–275, 1993.

[3] G. Benson. An Algorithm for Finding Tandem Repeats of Unspecified Pattern Size. In *RECOMB'98*, pages 20–29. ACM-, 1998. Proc.RECOMB'98,New-York.

[4] J.D. Biggins. A note on repeated sequences in Markov chains. *Advances in Applied Probability*, 19:739–742, 1987.

[5] M.Y. Borodovsky and J. Kleffe. First and second moments of counts of words in random texts generated by markov chains. *CABIOS*, 8:433–441, 1992.

[6] S. Breen, M.S. Waterman, and N. Zhang. Renewal theory for several patterns. *J. Appl. Prob.*, 22:228–234, 1985.

[7] V. Brendel, J.S. Beckmann, and E.N. Trifonov. Linguistics of nucleotide sequences: Morphology and comparison of vocabularies. *J. Biomol. Struct. Dyn.*, 4:11–21, 1986.

[8] P. Bucher, K. Karplus, N. Moeri, and K. Hoffman. A flexible motif search technique based on generalized profiles. *Computers Chem.*, 18(3):3–23, 1996.

[9] J.M. Claverie. Some useful statistical properties of position-weight matrice. *Computers and Chemistry*, 18(3):287–294, 1994.

[10] E. Coward. Word occurrence probabilities and repetetive regions in DNA sequences. In *Proc. MABS'97, Rouen, August97*, 1997.

[11] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. John Wiley, 1968.

[12] J.W. Fickett. The Gene Identification Problem: an Overview for Developers. *Computers Chem.*, 20(1):103–118, 1996.

[13] Ph. Flajolet and R. Sedgewick. *Analysis of Algorithms*. Addison-Wesley, 1995.

[14] M.S. Gelfand. Prediction of Function in DNA Sequence Analysis. *Journal of Computational Biology*, 2:87–117, 1995.

[15] J.F. Gentleman and R.C. Mullin. The Distribution of the Frequency of Occurrence of Nucleotide Subsequences, Based on their Overlap Capability. *Biometrics*, 45:35–52, 1989.

[16] M. Geske, A. Godbole, A. Schafner, A. Skolnick, and G. Wallstrom. Compound Poisson Approximations for Word Patterns Under Markovian Hypotheses. *J. Appl. Prob.*, 32:877–892, 1995.

[17] L. Guibas and A.M. Odlyzko. String Overlaps, Pattern Matching and Nontransitive Games. *Journal of Combinatorial Theory,* Series A, 30:183–208, 1981.

[18] A.K. Konopka and G.W. Smythers. Distan-a Program which Detects Significant Distances between Short Oligonucleotides. *Comput. Appl. Biosci.*, 3, 1987.

[19] S. Kurtz and G. Myers. Estimating the Probability of Approximate Matches. In *CPM'97*, Lecture Notes in Computer Science. Springer-Verlag, 1997.

[20] S.R. Li. A Martingale Approach to the Study of Occurrences of Sequence Patterns in Repeated Experiments. *Ann. Prob.*, 8:1171–1176, 1980.

[21] W. Li. The study of correlation structures of DNA sequences: a critical review. *Computers Chem.*, 21(4):257–271, 1997.

[22] R. Lundstrom. *Stochastic Models and Statistical Methods for DNA Sequence Data.* Phdthesis, University of Utah, 1990.

[23] G. Mengeritzky and T.F. Smith. Recognition of Characteristic Patterns in Sets of Functionally Equivalent DNA Sequences. *Comput. Appl. Biosci.*, 3:223–227, 1987.

[24] Pierre Nicodème, Bruno Salvy, and Philippe Flajolet. Motif statistics. In *ESA'99*, volume 1643 of *Lecture Notes in Computer Science*, pages 194–211. Springer-Verlag, 1999. Proc. European Symposium on Algorithms-ESA'99, Prague.

[25] P.A. Pevzner, M. Borodovski, and A. Mironov. Linguistic of Nucleotide sequences:The Significance of Deviations from the Mean: Statistical Characteristics and Prediction of the Frequency of Occurrences of Words. *J. Biomol. Struct. Dynam.*, 6:1013–1026, 1991.

[26] B. Prum, F. Rodolphe, and E. de Turckheim. Finding Words with Unexpected Frequencies in DNA sequences. *J. R. Statist. Soc. B.*, 57:205–220, 1995.

[27] M. Régnier. Efficient Computation of Unusual Words Expectation, 1999. presented at WORDS'99.

[28] M. Régnier and W. Szpankowski. On Pattern Frequency Occurrences in a Markovian Sequence. *Algorithmica*, 22(4):631–649, 1997. preliminary draft at ISIT'97.

[29] M. Régnier and W. Szpankowski. On the Approximate Pattern Occurrences in a Text. In IEEE Computer Society, editor, *Compression and Complexity of SEQUENCES 1997*, pages 253–264, 1997. In Proceedings SEQUENCE'97,Positano, Italy.

[30] S. Schbath. *Etude Asymptotique du Nombre d'Occurrences d'un mot dans une Chaine de Markov et Application à la Recherche de Mots de Frequence Exceptionnelle dans les Sequences d'ADN.* Thèse de 3e cycle, Université de Paris V, 1995.

[31] M.J. Shulman, C.M. Steinberg, and N. Westmoreland. The Coding Function of Nucleotide Sequences can be Discerned by Statistical Analysis. *J. Theor. Biol.*, 88:409–420, 1981.

[32] M.S. Tanushev and R. Arratia. Central Limit Theorem for Renewal Theory for Several Patterns. *Journal of Computational Biology*, 4(1):35–44, 1997.

[33] M. Waterman. *Introduction to Computational Biology.* Chapman and Hall, London, 1995.

[34] J.c. Wootton and S. Federhen. Analysis of Compositionally Biased Regions in Sequence Databases. In R. F. Doolittle, editor, *Computer Methods for Macromolecular Sequence Analysis*, volume 266 of *Methods in Enzymology*, pages 554–571. Academic Press, 1996.